

陆云飞

算法工程师

北京

+86 17090149380

✉ Earthson.Lu@gmail.com

📖 blog.earthson.net



大数据和数据挖掘技术全栈工程师

求职意向

人工智能 机器学习, 数据挖掘, 推荐系统等相关领域的工作

量化交易 希望能够从事基于机器学习方案的程序化交易系统的构建

主要技能

Scala 主语言, 函数式爱好者, 熟悉大部分特性, 熟练编写应用和类库

Python 主语言, 主要于机器学习和数据挖掘, 熟练 Scikit-Learn/Pandas, 擅长爬虫

大数据 熟悉大数据平台, 深刻理解 Spark 的原理, 擅长 Spark 应用及调优

机器学习 熟悉基本原理, 实现过部分算法, 拥有丰富项目经验

全栈开发 熟悉 C/C++, Java, JavaScript 等, 熟练使用 Linux, 全栈开发工程师

社区贡献 Spark, XGBoost

教育经历

2009–2013 本科, 吉林大学, 计算机科学与技术.

ACM/ICPC Regional 2010 Harbin Gold

工作经历

13.06–13.12 工程师, 秒针系统, 北京.

- 基于 Storm 流式框架完成对公司广告投放系统的日志实时监控。架构和开发全栈使用 inotify 监控新日志, redis 作为日志队列, 用 Storm 消费队列。利用时间窗自动分段汇总, 完成对 MySQL 数据库的更新。了毫秒级的任意时间区间汇总数据查询
- 基于 Python coroutine 的电商爬虫系统。架构和开发全栈异步非阻塞爬虫服务, tproxy 作代理池, MongoDB 作数据仓库

14.01-至今 算法工程师, 明略数据, 北京.

- DataInsight 大数据挖掘平台。作为架构和主开发
 - 架构。使用 Spark 作为数据和计算引擎, 使用 Akka 和 Play 作为产品框架, 实现 DAG 实时任务调度引擎
 - 开发。3w+ 的 Scala 开发, 算法以及前端全栈开发
 - 运维。使用 Anaconda 作为基础平台, 提供 Python 和 R 的环境。Spark 的维护和调优。集成测试环境维护
 - 管理。代码 review, 开发任务分配
- 银联商务店铺和大众点评商铺匹配
使用经纬度, 地址, 店名等信息, 提取最大公共前缀, 子串和子序列等, 利用 Logistic Regression 分类算法完成商户的匹配。结果 F1-score 达到了接近 90%。因为数据量较大 (数百万商户), 利用 Spark 平台, 先使用 GeoHash 先对候选集进行粗筛, 再用 Logistic Regression 训练模型, 对结果进行排序, 取匹配度最好的分店作为候选。
- 华泰证券用户推荐 POC。基于用户的历史交易记录, 向用户推荐产品类别
使用用户的历史交易记录按时间窗分段分别汇总用户和产品的特征, 使用 XGBoost 训练模型预测用户在下一个时间段的购买行为并按分值排序, 取 topK 作为推荐结果。在测试集上获得了 70% 的 F1-Score
- 江苏银行舆情分析 POC 爬虫系统
使用 Python 的 asyncio 系统设计了一套异步非阻塞的爬虫系统。拥有自动随机 ip 代理, host 流量均衡, 增量更新等特性。耗时两周
- 银联商务 MCC 清洗
使用用户的历史刷卡行为, 分析汇总得到 mcc 和商铺的刷卡行为分布, 并汇总商铺的名称等文本特征, 使用 XGBoost 训练分类模型, 最后在餐饮类别上获得了 90+% 的 F1-Score

个人爱好

计算机 热衷于钻研新技术, 函数式编程爱好者

羽毛球 很好的陪练

语言

Engilsh 大学英语四级